Module No.1

Correlation and Regression

Correlation

In many practical situations more than two variables are involved in a study which are interrelated.

In case of health study we need to consider (i) weight, height(ii) cholesterol, blood pressures etc.

To study performance of a student in an examination the factors to be considered could be number of hours studied, number of lectures attended, etc.

In a study when two variables (X, Y) are considered then the corresponding data is termed as bivariate data.

Correlation and regression analysis are the most commonly used techniques for studying the relationship between two quantitative variables

Correlation is concerned with the investigation of **two** variables usually measured on same item and are **logically related**.

While studying these Bivariate data we are interested to know:

- Whether a relationship exists between those variables;
- if so, how strong that relationship is;
- What form that relationship takes;
- Can we make use of that relationship for predictive purposes?

Correlation describes the **strength of the relationship**.
But it is not concerned with `cause' and 'effect'.
We can explore the relationship between two quantitative variables
- Graphically, by constructing a scatter plot.
- Numerically, by constructing correlation coefficient

Scatter plot :

The observation pairs (X,Y) are plotted on X- axis and Y- axis which is called as Scatter plot. It gives a visual impression of how two values are related. The relationship between two variables is called their **correlation** .The scatter plot gives us an idea about the relationship between X and Y in the form of

Shape : whether it is linear or nonlinear (i.e. curved)

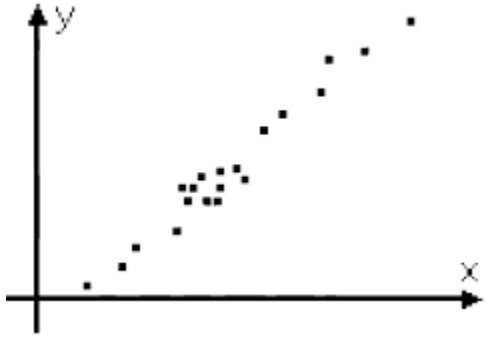
Direction : positive or negative slope

Strength : how tight or spread out the points are

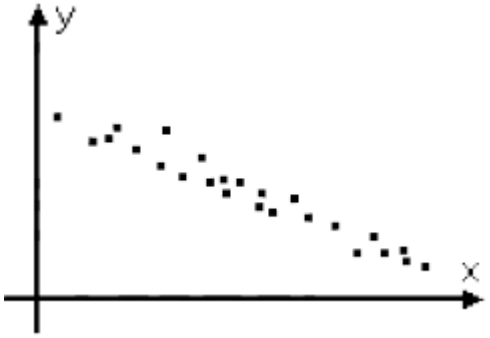
Correlations may be positive (rising), negative (falling), or null (uncorrelated).

i) A positive *linear relationship* means that as *X* increases, *Y* also increases. If the pattern of dots, slopes from lower left to upper right, it indicates a positive correlation between the variables being studied (i.e. a positive slope.)

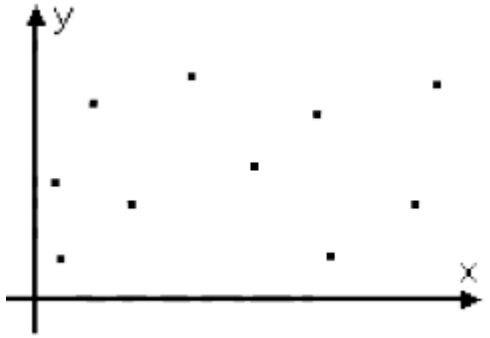
Eg. Height and weight



ii) A *negative linear relationship* means that as X increases, Y decreases. If the pattern of dots, slopes from upper left to lower right, it indicates a negative correlation. (i.e. a negative slope.) Eg. Price and demand



iii*) Null* means no relationship or a curved relationship.
Eg. I.Q. and shirt size of a student

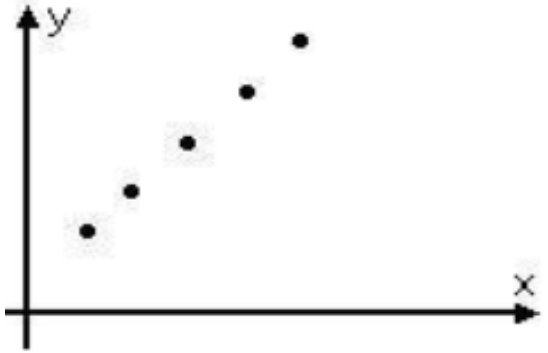


Interpretation of relationship from scatter diagram:

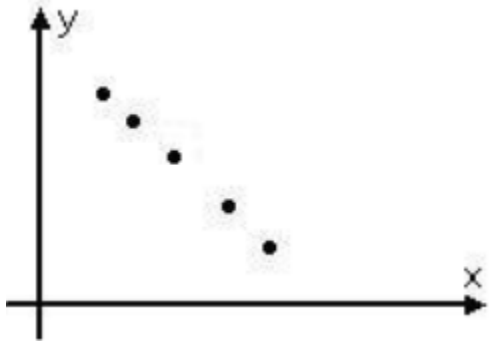
When the plotted data points come closer and closer, to make a straight line, the correlation between the two variables becomes higher or relation becomes stronger. The relation may be a) perfect b) strong c) Weak d) null e) non linear.

(i)If values of X increases (decreases) then corresponding values of Y increases (decreases) , data points fall on a straight line, then the variables are having **perfect positive correlation** .

(ii)If values of X increases (decreases) then corresponding values of Y decreases (increases) , data points fall on a straight line, then the variables are having **perfect negative correlation**



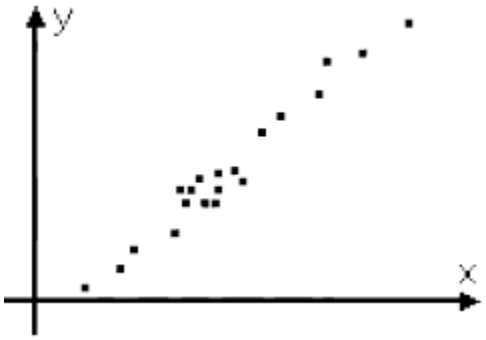
perfect positive correlation



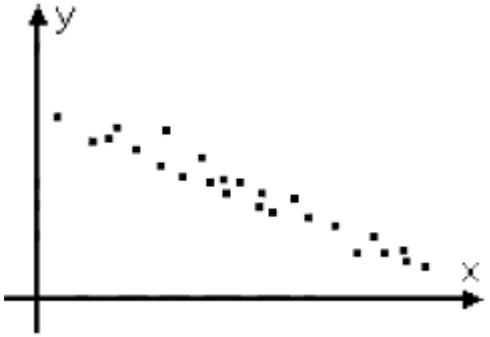
perfect negative correlation

(iii)If values of X increases (decreases) then corresponding values of Y increases (decreases) , data points do not fall on a straight line, then the variables are having **strong positive correlation** .

(iv)If values of X increases (decreases) then corresponding values of Y decreases (increases) , data points do not fall on a straight line, then the variables are having **strong negative correlation**

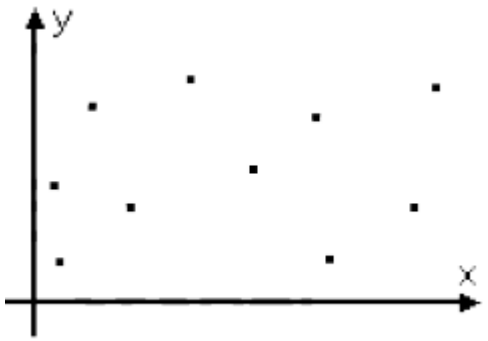


strong positive correlation

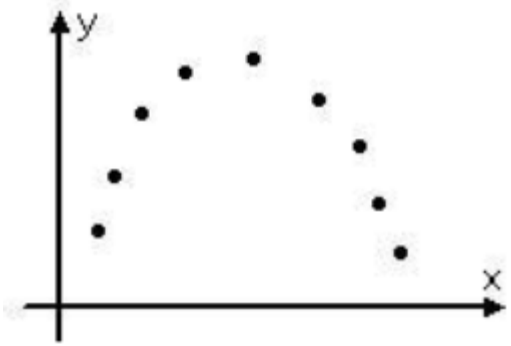


strong negative correlation

(v)The points are just too scattered without following any pattern like in Fig then there is no relation between the variables.



(vi)Data points are showing a perfect pattern, but it is non linear. All the points are falling on some curve which can be visualized, like in Fig, and then the variables have non linear relation.



Merits of Scatter plot

1. It is easy to plot.
2. It is very easy and simple to understand even at a single glance.
3. Abnormal values in a sample can be easily detected.
4. Scatter plot has an ability to show nonlinear relationships between variables.

Demerits of Scatter plot

1. It does not give a numerical measure of correlation.
2. No mathematical or algebraic treatment is possible to the result.
3. The method is useful only when number of terms is small.
4. It cannot be applied to qualitative data.

To overcome the limitation of scatter diagram, we need to have numerical measure of correlation.

We shall study two types numerical measure of correlation

(i) Karl Pearson's correlation coefficient (r)

(ii) Rank correlation coefficient (R)
(Spearman's rank correlation coefficient)

Karl Pearson's correlation coefficient (r) :

Karl Pearson introduced the correlation coefficient which quantifies the strength and direction of the **linear** relationship between two variables measured on interval. It is denoted by 'r'.

Let X and Y be two variables and number of observations be n , Karl Pearson's correlation coefficient r between them is given by

$$r = cor(X,Y) = r_{xy} = \frac{cov(X,Y)}{sd(X)sd(Y)} \qquad \text{where}$$

$$cov(X,Y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{n},\ sd(X) = \sqrt{\frac{\sum(x-\bar{x})^2}{n}},\ sd(Y) = \sqrt{\frac{\sum(y-\bar{y})^2}{n}}$$

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} = \frac{\sum(x')(y')}{\sqrt{\sum(x')^2\sum(y')^2}}$$

Basic Concepts

STANDARD DEVIATION

Standard deviation is defined as the square root of the mean of the square of the deviation from the arithmetic mean.

$$S.D. = \sigma = \sqrt{\frac{\Sigma f (x - \bar{x})^2}{\Sigma f}}$$

Note. 1. The square of the standard deviation σ^2 is called variance.

2. σ^2 is called the second moment about the mean and is denoted by μ_2.

Basic Concepts

STANDARD DEVIATION

Standard deviation is defined as the square root of the mean of the square of the deviation from the arithmetic mean.

$$S.D. = \sigma = \sqrt{\frac{\Sigma f (x - \bar{x})^2}{\Sigma f}}$$

Note. 1. The square of the standard deviation σ^2 is called variance.

2. σ^2 is called the second moment about the mean and is denoted by μ_2.

Covariance

covariance is a measure of the relationship between two random variables. The metric evaluates how much – to what extent – the variables change together.

$$cov(X, Y) = \frac{\sum(x - \bar{x})\,(y - \bar{y})}{n}$$

Basic Formulae

Expectation=$E(x) = \frac{\sum x}{n} = \bar{x}$

Standard deviation = sqrt(Variance)

Variance= $V(x) = \frac{\sum(x-\bar{x})^2}{n} = E(x^2) - [E(x)]^2$

$$cov(X,Y) = \frac{\sum(x-\bar{x})\,(y-\bar{y})}{n} = E(xy) - E(x)E(y)$$

Properties of *r*:

(i)The value of *r* always lies between -1 and +1
i.e.$-1 \leq r \leq 1$

(ii) The correlation coefficient *r* is not affected by shift of origin . i.e. r(x,y)=r(x-a , y-b)

(iii) The correlation coefficient *r* is not affected by change of scale i.e. r(x,y)= r(x/h , y/k)

Note :

(i) When r = 1, ⇒ there is perfect positive linear relationship so that all the points in a scatter plot of the data lie exactly on a straight line with a positive slope.

(ii) When *r* = -1,⇒ there is perfect negative linear relationship so that all the points in a scatter plot lie exactly on a straight line with a negative slope

(iii) Correlation coefficient is zero does not mean that the variables are not associated. It simply indicates that there is no *linear* relationship between the variables.

Correlation coefficient is zero does not mean that the variables are not associated. It simply indicates that there is no *linear* relationship between the variables The variables may have quadratic or any other non linear relationship among the variables. E.g. Y = X^2.i.e variables may be dependent. But if variables are independent then there does not exist any kind of relationship, neither linear nor non linear, hence correlation coefficient becomes zero. So that, if variables are independent, correlation coefficient is zero but converse is not true. i.e. if correlation coefficient is zero, it does not imply that variables are independent

Formula of r

(i)When all five sums are given $\sum x, \sum y, \sum x^2, \sum y^2$ and $\sum xy$.

$$r = \frac{\sum(xy) - \frac{(\sum x)(\sum y)}{n}}{\sqrt{(\sum(x)^2 - \frac{(\sum x)^2}{n})(\sum(y)^2 - \frac{(\sum y)^2}{n})}}$$

(ii) If $\bar{x}$, $\bar{y}$ are integers

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} = \frac{\sum(x')(y')}{\sqrt{\sum(x')^2\sum(y')^2}}$$

(iii)If $\bar{x}$, $\bar{y}$ are not integers, put $d_x = x - A$, $d_y = y - B$

$$r = \frac{\sum(d_x d_y) - \frac{(\sum d_x)(\sum d_y)}{n}}{\sqrt{(\sum(d_x)^2 - \frac{(\sum d_x)^2}{n})(\sum(d_y)^2 - \frac{(\sum d_y)^2}{n})}}$$

Example 1 : Calculate the correlation coefficient from the following data

x	229	226	228	227	230	232	223	225	232	228
y	264	261	260	259	263	260	264	266	256	257

Example 1 : Calculate the correlation coefficient from the following data

x	229	226	228	227	230	232	223	225	232	228
y	264	261	260	259	263	260	264	266	256	257

$\bar{x}$ is 228, $\bar{y}$ is 261

Example 1 : Calculate the correlation coefficient from the following data

x	229	226	228	227	230	232	223	225	232	228
y	264	261	260	259	263	260	264	266	256	257

$\bar{x}$ is 228, $\bar{y}$ is 261

Use the formula

r= $: \dfrac{\Sigma(x')(y')}{\sqrt{\Sigma(x')^2 \Sigma(y')^2}}$

sr no	x	y	$x' = (x - \bar{x})$	$y' = (y - \bar{y})$	$x'y'$	$(x')^2$	$(y')^2$
1	229	264	1	3	3	1	9
2	226	261	-2	0	0	4	0
3	228	260	0	-1	0	0	1
4	227	259	-1	-2	2	1	4
5	230	263	2	2	4	4	4
6	232	260	4	-1	-4	16	1
7	223	264	-5	3	-15	25	9
8	225	266	-3	5	-15	9	25
9	232	256	4	-5	-20	16	25
10	228	257	0	-4	0	0	16
	2280	2610	0	0	-45	76	94

r= $: \dfrac{\Sigma(x')(y')}{\sqrt{\Sigma(x')^2 \Sigma(y')^2}}$ =-0.5324

Example 2 : Calculate the correlation coefficient from the following data

X	100	110	120	130	140	150	160	170	180	190
Y	45	51	54	61	66	70	74	78	85	89

Example 2 : Calculate the correlation coefficient from the following data

X	100	110	120	130	140	150	160	170	180	190
Y	45	51	54	61	66	70	74	78	85	89

$\bar{x}$ is 145, $\bar{y}$ is 67.3

Example 2 : Calculate the correlation coefficient from the following data

X	100	110	120	130	140	150	160	170	180	190
Y	45	51	54	61	66	70	74	78	85	89

$\bar{x}$ is 143, $\bar{y}$ is 67.3

Put A= 143 , B= 68

Use the formula

$$r= \frac{\Sigma(d_x d_y) - \frac{(\Sigma d_x)(\Sigma d_y)}{n}}{\sqrt{(\Sigma(d_x)^2 - \frac{(\Sigma d_x)^2}{n})(\Sigma(d_y)^2 - \frac{(\Sigma d_y)^2}{n})}}$$

sr no	x	y	$d_x = x - A$	$d_y = y - B$	$d_x d_y$	$(d_x)^2$	$(d_y)^2$
1	100	45	-45	-23	1035	2025	529
2	110	51	-35	-17	595	1225	289
3	120	54	-25	-14	350	625	196
4	130	61	-15	-7	105	225	49
5	140	66	-5	-2	10	25	4
6	150	70	5	2	10	25	4
7	160	74	15	6	90	225	36
8	170	78	25	10	250	625	100
9	180	85	35	17	595	1225	289
10	190	89	45	21	945	2025	441
	1450	673	0	-7	3985	8250	1937

$$r = \frac{\sum(d_x d_y) - \frac{(\sum d_x)(\sum d_y)}{n}}{\sqrt{(\sum(d_x)^2 - \frac{(\sum d_x)^2}{n})(\sum(d_y)^2 - \frac{(\sum d_y)^2}{n})}} = 0.998129$$

Exercise

Example 1: *Ten students got the following percentage of marks in Economics and Statistics.*

Roll No.	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Marks in Economics	*78*	*36*	*98*	*25*	*75*	*82*	*90*	*62*	*65*	*39*
Marks in Statistics	*84*	*51*	*91*	*60*	*68*	*62*	*86*	*58*	*53*	*47*

Calculate the coefficient of correlation.

1) r=0.78

Example 2 . *In a study between the amount of rainfall and the quantity of air pollution removed the following data were collected.*

Daily Rainfall in 0.01 cm	*4.3*	*4.5*	*5.9*	*5.6*	*6.1*	*5.2*	*3.8*	*2.1*
Pollution Removed (mg/m^3)	*12.6*	*12.1*	*11.6*	*11.8*	*11.4*	*11.8*	*13.2*	*14.1*

Calculate the coefficient of correlation.

Misc. example

Ex 3

For 10 pairs of values of x and y the following values are determined: Later on it was found that one pair of values was taken as (34, 47) instead of (43, 74).Determine the correct value of the coefficient of correlation if Mean(X) = 30.1, Mean(Y) = 47.8, S.D.(X)=6.2, S.D.(Y)=9.5, r= 0.72

Exercise 4

In two sets of variables X and Y with 50 observations each, the following data were observed :

$$\bar{X} = 10, \sigma_X = 3, \bar{Y} = 6, \sigma_Y = 2 \text{ and } r(X, Y) = 0{\cdot}3$$

But on subsequent verification it was found that one value of X (= 10) and one value of Y (= 6) were inaccurate and hence weeded out. With the remaining 49 pairs of values, how is the original value of r affected ?